一种基于多尺度神经网络的船舶目标检测方法

吴园园[2†]，臧柏涵[3†]，王兴华[1*]，刘帅[2]，张宗良[3]

（[1]集美大学航海学院，福建 厦门 361021；
[2]中国船舶重工集团公司第七〇七研究所，天津 300131；
[3]集美大学计算机工程学院，福建 厦门 361021）

†共同第一作者，*通讯作者

摘要：近年来，船舶智能化的发展对船舶目标的检测与分类精度要求越来越高，准确检测并识别船舶的类型及判断船舶的位置是船舶安全航行重要保障。由于船舶目标光学成像过程中易受到风、流、雨、雾等外部背景环境影响，导致基于深度学习的船舶目标检测算法性能降低；同时，船舶类型多样、形态各异以及几何尺寸大小不一等因素均使得船舶目标的检测和识别存在一定的困难。鉴于此，本文提出一种基于多尺度神经网络的目标检测方法用以提高光学影像中船舶目标的检测精度。该方法采用卷积神经网络（Convolutional Neural Networks）对图像予以特征提取，通过改进的基于 CSPDarkNet 骨干网络以及多尺度网络以实现船载光学摄像头对水上船舶目标的准确检测，提高模型对小目标和密集目标的检测精度。同时利用标签平滑化来防止模型陷入过拟合，并采用非极大值抑制降低重复检测。实验结果表明本文所提出的方法在 Ship-Detection 数据集上均值的平均精度(Mean Average Precision, mAP)可达 84.80%，与 Faster-RCNN、CO-DETR 等先前目标检测的研究方法相比，检测效果更好，更具备潜在的应用优势。

关键词：船舶目标检测；深度学习；神经网络；多尺度神经网络

中图分类号：U　**文献标志码**：A

A multi-scale network-based approach for optical imagery ship detections

WU Yuanyuan[2†], ZANG Bohan[3†], WANG Xinghua[1*], LIU Shuai[2], ZHANG Zongliang[3]

(1. Navigation College, Jimei University, Xiamen 361021, Fujian, China;
2. CSIC 707 Institute, Tianjin 300131, China;
3. College of Computer Engineering, Jimei University, Xiamen 361021, Fujian, China)

Abstract：In recent years, there has been an increasing demand for higher detection and classification accuracy of ship targets to enable safe ship navigation, driving the development of ship intelligence. However, the performance of deep learning-based ship target detection algorithms is affected by the optical imaging process of ship targets, which can be easily disrupted by environmental factors such as wind, current, rain, and fog. Additionally, the diverse range of ship types, morphologies, and sizes pose challenges for accurate detection and identification of ship targets. To address these challenges, this paper proposes a multi-scale neural network-based target detection method for improving the accuracy of ship target detection in optical images. The proposed method employs a Convolutional Neural Networks (CNN) to extract image features. The improved backbone of CSPDarkNet and multi-scale network is used to realize the accurate detection of the ship-borne optical camera

on the water ship target, and the detection accuracy of the model for small targets and dense targets is improved. Furthermore, label smoothing to prevent overfitting, and non-maximum suppression to reduce repetitive detections. Experimental results demonstrate that the proposed model achieves accurate detection of ship targets on water and can be used for the detection of small and intensive targets. The mean average precision (mAP) of the proposed method on the Ship-Detection dataset reaches 84.80, which outperforms previous research methods such as Faster-RCNN, DINO and offers greater potential for practical applications.
Key words: ship detections; deep learning; neural network; multi-scale neural network

0 引 言

伴随着水路运输和海洋经济的快速发展，推动船舶智能化进程的发展是当前社会发展的迫切需要。船舶作为重要的运输载体和军事目标，对船舶实现准确的检测、分类和识别，有助于改善航行安全及提高船员工作效率，同时在海事交通监管、维护国家海洋权益和海洋安全等领域具有重要的应用价值和战略意义。†

传统基于光学影像的船舶目标检测方法主要有三种：基于影像角点、基于影像边缘特征的和基于影像区域特征的目标检测方法。在基于影像角点的目标检测上，Smith [1]提出了 SUSAN 算法，通过在图像核心点设定一个大小为 37 个像素的圆形模板，计算模板中与核心点相似亮度值的个数，并对初始设定角点采用非极大值抑制求得最后的角点；基于边缘特征的目标检测通常则是对输入的图像予以平滑化处理，用以减少或消除图像中噪声影响，并进一步采用边缘检测算子（Roberts[2]、Canny[3]等）得到图像边界点；基于区域特征的目标检测方法则主要采用图像边缘与区域相结合的处理方式，并通过连通的灰度二值化图和多阈值处理来实现目标的检测。

近年来，随着大数据应用的日趋深入和计算处理速度的不断提高，基于卷积神经网络[4]（Convolutional Neural Network, CNN）的智能化目标检测方法取得了极大进步。目前，基于 CNN 的目标检测方法主要分为 One-Stage 和 Two-Stage 两种策略[5]；其中，One-Stage 是指直接对输入的图像进行目标检测，而 Two-Stage 则是在 One-Stage 的基础上加入区域生成网络（Region Proposal Network, RPN）网络[6]，通过引入区域生成网络(RPN)对待检测目标进行位置约束。在基于 CNN 进行目标检测的研究中，Alexander 等先后提出了 R-CNN、Fast R-CNN 和 Faster R-CNN[6]等一系列经典神经网络结构，这些网络结构和算法为智能化的光学影像目标检测奠定了基础。神经网络是一种端到端模型，避免了人工设计特征的巨大消耗。随着神经网络模型的不断深化，图像目标检测的准确率也不断提高。

但由于上述 R-CNN 系列算法自身结构原因，传统的 R-CNN 网络模型对小目标的检测精度较低，难以满足部分应用场景中的要求。针对该问题，本文提出一种基于多尺度的神经网络船舶检测算法。Sermanet 开创性的提出多尺度概念，其可以很好的解决在在影像中不同尺度大小的特征目标对算法性能的影响，同时可以提高模型的鲁棒性，本文提出的一种基于多尺度的神经网络模型算法具有较高的准确率和良好的检测速率，能够满足船舶目标检测对精度和实时性的要求。

综合上述分析，同时考虑到船载传感器得到的光学数据存在不同大小（所占的像素区域）的船舶目标，本文开展基于多尺度的船舶目标检测算法研究。本文主要分析多尺度检测网络的网络结构和损失函数，制作船舶目标检测所需的数据集，并针对船舶特性进行改进和优化，同时进行多轮不同的实验进行对比，验证了算法的检测效果和性能。

1　基于多尺度的船舶目标检测网络

我们的网络模型结构分为输入层、骨干网络、颈部网络以及输出层四个部分。整体的网络模型是一种 One-Stage 的目标检测，如图 1 所示。在骨干网络中，我们采用基于 CSPDarkNet 的基本框架，并使用更强大的基本构建块（见 1.1）来提升模型的准确性，并且据此来调整颈部网络中模型的深度、宽度和分辨率等参数（见 1.2）。在开始训练之前的输入层中使用 Mosaic 方法进行数据增强（见 1.3）。与此同时，我们采用标签平滑方法作为正则化方法，同时使用 Focal-Loss[12]和 GIoU[13]作为损失函数来优化模型（见 1.4）。整体的网络模型算法结构图如图 1 所示。

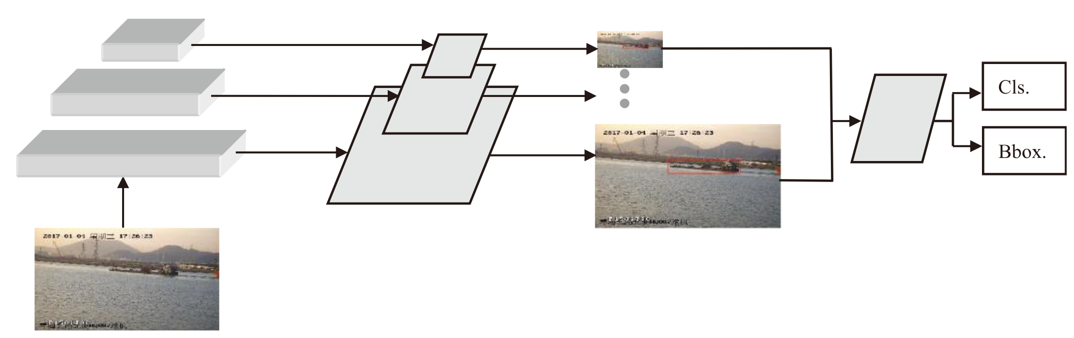


图 1 基于多尺度的船舶目标检测网络模型算法结构图

1.1 改进的骨干网络

我们采用改进的 CSPDarkNet[16]作为骨干网络，传统的基本构建块如 2.a 所示，由 1×1 和 3×3 两层卷积组成。考虑到船舶有时候的分布较为密集，并且在目标检测中较大的有效感受野对于密集检测任务更有效[18]，在其基本构建块中保留 3×3 的卷积并且采用 5×5 的深度卷积，从而增加有效感受野，如 2.b 所示。由于改进后的基本构建块中增加了卷积层的个数，这会导致检测速度的降低，所以我们减少了基本构建块的使用量，并且对整体的网络做出了一些修改，使得最终的效果最好。我们所使用的骨干网络整体结构图如图 所示。

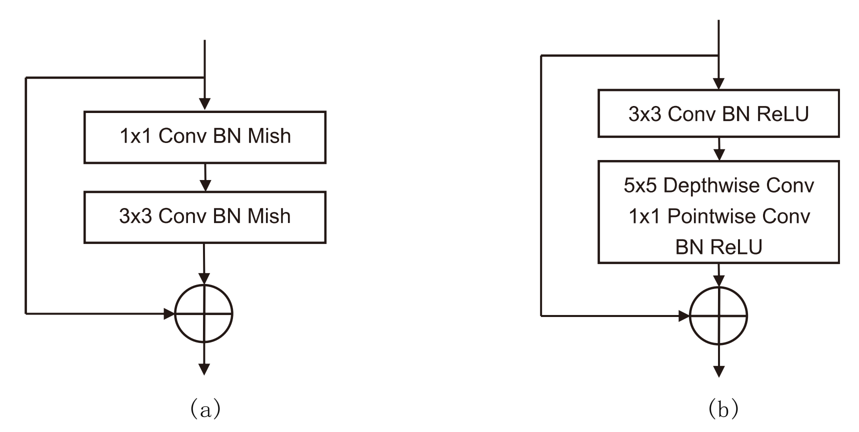


图 2 不同骨干网络中的基本构建块。(a) CSPDarkNet 中所采用的基本构建块。由 1×1 和 3×3 两层卷积组成。(b) 我们所采用的改进的基本构建块。通过引入 5x5 的 Depthwise 卷积从而增加有效感受野，并且通过该卷积降低运算成本。

此外， 图 中每个卷积层后需要跟一个 BatchNorm[10]和一个 ReLU[11]激活函数，其计算方式如下：

$$F(X) = ReLU\Big(BN\big(Conv2d(X)\big)\Big) \tag{1}$$

其中 BN 表示批量归一化，虽然在输入层已经对数据进行归一化的预处理，但是对于深层神经网络来说，在训练过程中参数的更新仍然会造成参数的剧烈变化，这种变化通常会影响到最终训练出来模型的效果，因此每次卷积运算之后需要在进行一次归一化。ReLU 激活函数表达形式如下：

$$G(X) = Max(0, X) \tag{2}$$

使用 ReLU 与其他激活函数相比计算更为简单[11]，并且可以使一部分的参数输出为 0，减少了参数之间的相互关系，有利于提升目标检测的速度。

Type	Filters	Size	Output
Convolutional	32	3×3/2	320×320
Convolutional	32	3×3	320×320
Convolutional	64	3×3	320×320
Convolutional	64	3×3/2	320×320
Convolutional	128	3×3	
Depthwise Convolutional	128	5×5	
Pointwise Convolutional	128	1×1	
Avgpool			160×160
Convolutional	256	3×3/2	160×160
Convolutional	256	3×3	
Depthwise Convolutional	256		
Pointwise Convolutional	256		
Avgpool			80×80
Convolutional	512	3×3/2	80×80
Convolutional	512	3×3	
Depthwise Convolutional	512		
Pointwise Convolutional	512		
Avgpool			40×40
Convolutional	1024	3×3/2	40×40
Maxpool		5×5	
Maxpool		5×5	
Maxpool		5×5	
Convolutional	1024	3×3	
Depthwise Convolutional	1024		
Pointwise Convolutional	1024		
Avgpool			20×20

图 3 我们所采用的骨干网络整体结构

1.2 颈部网络

对于目标检测任务来说，多尺度特征金字塔是必不可少的。所谓的多尺度是对信号的不同颗粒进行采集，利用不同尺度下能够观察到不同特征，从而完成不同尺寸的检测任务。本文船舶目标检测任务中的颈部网络在骨干网络提取的特征基础上进一步特征融合，帮助网络感知不同尺度上的目标，并提供更多上下文信息。

为了适应骨干网络的改变，并且考虑到训练速度因素，我们在颈部网络中拓展基本构建块，将更多的计算放在颈部网络完成，从而在速度与精度上获得更好的权衡。

1.3 数据增强

我们使用 Mosaic 方法处理数据集，其主要思想是将四张图片进行随机裁剪，再拼接到一张图片上进行训练。这样处理数据不仅可以增加数据的多样性，使用比图像个数多的原图进行训练，还能增强模型的鲁棒性，让模型具有泛化能力。

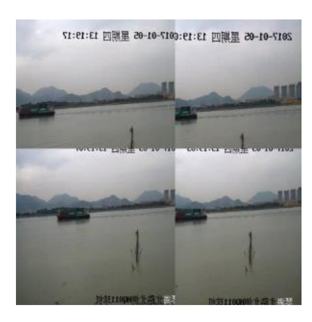



图 4 使用 Mosaic 方法拼接的船舶检测图像

1.4 损失函数

由于图像中会含有不同类别的船舶，同时存在类别非常少的数据，如帆船含量极少，因此本文针对这类问题本文采用 Focal Loss[12] 作为本文的损失函数，其具体的计算方式如下：

$$\begin{cases} -a(1-p)^{r}\,log_{10}(p) & if\ y=1 \\ -(1-a)p^{r}\,log_{10}(1-p) & if\ y=0 \end{cases} \tag{3}$$

其中，r为 2，α为 0.25。Focal Loss 关注大样本数据，较其他的损失函数仅针对存在预测标注的样本数据进行约束，导致输出图像边界的数据无法得到约束，当同一训练样本中存在大量无标注数据时会出现梯度爆炸的情况。

同时针对算法预测的定位框作为损失函数的约束对象，即 GIoU Loss[13]，使模型预测出的边界与专家标注的边界之间度量距离降低，其具体计算方式如下：

$$IoU\ Loss = -\,log_{10}\frac{Intersection\left(\hat{y},y\right)}{Un\left(\hat{y},y\right)} \tag{4}$$

$$GIoU = IoU - \frac{C\backslash Intersection(\hat{y},y)}{C} \tag{5}$$

其中，$Intersection$表示两个定位框之间的交集，Un表示两个定位框之间的并集，C为最小的定位框在图像中所占的面积。可以得到两个定位框之间的交并比，而以两个定位框之间的交并比为约束，可以降低预测与实际物体之间的度量距离。

2 多尺度模型检测模型优化

2.1 非极大值抑制

在目标检测任务中，通常会生成大量的定位框，可能同一例实例中会存在大量位置相似重叠框，采用非极大值抑制，从一系列重叠框中选择出最佳的边界框，仅保留一定范围内概率最大的边界框，其具体计算方式如下：

$$boxlist = \left[\hat{y}_1, \hat{y}_2 \dots \hat{y}_n\right] \tag{6}$$

$$s_i = \begin{cases} s_i & iou(M, b_i) < N_t \\ s_i(1 - iou(M, b_i)) & iou(M, b_i) \geq N_t \end{cases} \tag{7}$$

当 IOU 值超过所设定的阈值（普遍设置为 0.5，目标检测中常设置为 0.7，仅供参考），即对超过阈值的框进行抑制，抑制的做法是将检测框的得分设置为 0，如此一轮过后，在剩下检测框中继续寻找得分最高的，再抑制大于 IOU 阈值的框，直到最后会保留几乎没有重叠的框。这样基本可以做到每个目标只剩下一个检测框。

2.2 标签平滑化

在传统的 one-hot 编码标签的深度学习网络训练过程中，其中各个类别的约束方向都希望无限趋近于优化目标/类别的 1，其中非目标/类别则约束至 0，即在最终网络输出的预测张量中目标类别的概率趋近于 1，使得模型对于正确标签和错误标签间的方差过大，但会降低模型的稳健性，出现过拟合或者梯度爆炸的情况，这时候对于采用类熵的损失函数就无法完全发挥其效率。因此采用标签平滑化的方法，对标签进行平滑。

$$P_i = \begin{cases} 1 & if(i = y) \\ 0 & if(i \neq y) \end{cases} \tag{8}$$

$$P_i = \begin{cases} (1 - \xi) & if(i = y) \\ \dfrac{\xi}{K - 1} & if(i \neq y) \end{cases} \tag{9}$$

3 实验与结果分析

3.1 数据集构建

在深度学习的数据集占着决定性的作用，决定模型最终的检测效果，同时还需要足够的样本使得网络能够充分学习检测目标的特征图像，因此构建的船舶数据集如表 1 所示.

表 1 总数据集的分布情况

名称	总数据集	训练集	测试集
样本个数/个	7000	5600	1400

3.2 模型训练

本文采用 PyTorch[14]作为本模型方法的深度学习框架，在训练过程中，使用 Step 学习率优化方法，每 20 个 epoch 学习率乘以 0.1，初始学习率为 0.0003 和 Adam[15]的优化器，训练的批次大小为 8，本模型的实验方法在 CUDA 11.6，PyTorch 版本为 1.12，进行 300 次迭代训练。

3.3 性能对比

对提出的基于多尺度的船舶目标检测算法与其他基于深度学习的目标检测方法进行对比试验，为了保持训练的一致性，我们采用相同的训练参数（学习率、优化器、权重衰减策略等），表 2 为本文

提出的模型算法与其他主流算法之间的对比结果，其中表中的数据仅仅具备相对参考，因为验证数据集的标注存在一定误差。采用 mean Average Precision（mAP）作为评估指标，计算方式如：

$$mAP = \frac{\sum \frac{TP}{TP + FP}}{n} \tag{10}$$

其中 TP 表示真阳性率（即预测为真实际为真），FP 为假阳性率，n 表示总样本个数。

表 2 本文提出的模型算法与其他主流算法之间的对比结果

模型方法	mAP@.5:.95	mAP@0.5	mAP@0.75
Faster-rcnn[6]	0.804	0.987	0.960
Sparse R-CNN[22]	0.811	0.976	0.945
CO-DETR[19]	0.733	0.959	0.882
DINO[20]	0.845	0.987	**0.976**
DDQ[21]	0.821	0.987	0.952
Ours	**0.848**	**0.989**	0.959

3.4 检测结果

对算法的预测结果进行评估结果，输出模型预测结果。左上为 Faster R-CNN 算法检测效果，中上为 CO-DETR 算法检测效果，右上为 Sparse R-CNN 算法检测效果，左下为 DDQ 算法检测效果，中下为 DINO 算法检测效果，右下为我们的算法检测效果。其中 Sparse R-CNN 和 DDQ 算法对小目标船舶影像目标不敏感，出现了漏检情况；Faster R-CNN 算法的候选框出现了重复，对于较为复杂的情况检测会出现误检，此外 Faster R-CNN 区域候选网络的搜索算法只能在 CPU 上进行运行，这使得在 CPU 较弱的机器上进行推理时效率较低；CO-DETR 算法效果虽然与本文提出的多尺度检测模型效果均能正确检测，但是 CO-DETR 算法对于重叠影像的效果较我们的算法效果欠佳，图中靠后的船舶 Ground Truth 只有 50.4，在训练欠佳的情况下同样会出现漏检误检的情况。本文针对上述推理速度、空间单元格限制、重叠目标等情况，进行优化设计，即减少区域候选网络的复杂度，对单元格限制替换使用非极大值抑制进行限制，采用多尺度的算法进行区域检测提高对重叠目标的检测精度。

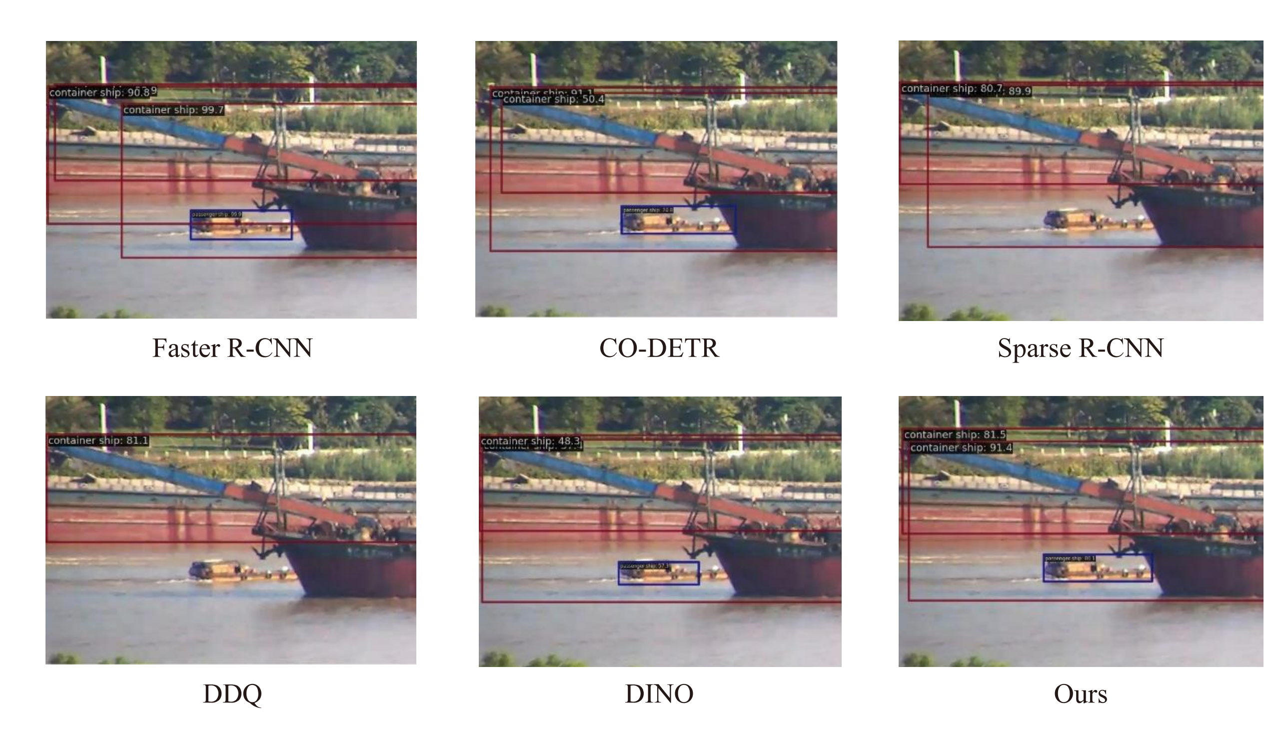


图 5 多尺度船舶检测模型与主流模型检测效果对比

4 结 论

本文基于多尺度目标检测算法模型，采用骨干网络与颈部网络提取出感兴趣区域，使用标签平滑化等优化策略对算法进行改进和优化，提出一种船舶目标实时检测的方法。同时，使用自制船舶图像数据集，通过设计好的模型在 Ubuntu 服务器进行训练，并与 Faster-RCNN、CO-DETR、Sparse R-CNN 等算法进行对比分析。实验结果表明：改进算法的 mAP 达到了 84.80%，优于其他算法，对于小目标以及重叠影像具有良好的检测效果。

参考文献

[1] Smith S M, Brady J M.SUSAN: A New Approach to Low Level Image Processing[J]. Int. Journal of Compuer Vision, 1997, 23(1):45-78.

[2] L. Roberts Machine Perception of 3-D Solids, Optical and Electro-optical Information Processing, MIT Press 1965

[3] Canny, J., A Computational Approach To Edge Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986.

[4] Fukushima,K. (2007). "Neocognitron". Scholarpedia.2(1): 1717. Bibcode: 2007SchpJ...2.1717F. doi:10.4249/scholarpedia.1717.

[5] Burke, D. L., & Ensor, J. (2017). Meta-Analysis Using Individual Participant Data: One-Stage and Two-Stage Approaches, and Why They May Differ. Tutorial in Biostatistics, 36(5), 855–875. doi:https://doi.org/10.1002/sim.7141.

[6] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. doi:10.48550/ARXIV.1506.01497.

[7] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection. doi:10.48550/ARXIV.1506.02640.

[8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. ECCV 2016. doi:10.1007/978-3-319-46448-0_2.

[9] Redmon, J., & Farhadi, A. (2016). YOLO9000: Better, Faster, Stronger. doi:10.48550/ARXIV.1612.08242.

[10] Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. doi:10.48550/ARXIV.1502.03167.

[11] Agarap, A. F. (2018). Deep Learning using Rectified Linear Units (ReLU). doi:10.48550/ARXIV.1803.08375.

[12] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[13] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. doi:10.48550/ARXIV.1902.09630.

[14] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

[15] Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. doi:10.48550/ARXIV.1412.6980.

[16] Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. doi:10.48550/ARXIV.2004.10934.

[17] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). CenterNet: Keypoint Triplets for Object Detection. doi:10.48550/ARXIV.1904.08189.

[18] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard S. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In NeurIPS, 2016.

[19] Zong, Zhuofan, Guanglu Song and Yu Liu. “DETRs with Collaborative Hybrid Assignments Training.” 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2022): 6725-6735.

[20] Zhang, Hao, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun-Juan Zhu, Lionel Ming-shuan Ni and Heung-yeung Shum. “DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection.” *ArXiv* abs/2203.03605 (2022): n. pag.

[21] Zhang, Shilong, Wang xinjiang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo and Kai Chen. “Dense Distinct Query for End-to-End Object Detection.” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023): 7329-7338.

[22] Sun, Pei, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang and Ping Luo. "Sparse R-CNN: End-to-End Object Detection with Learnable Proposals." *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020): 14449-14458.